DOCUMENT RESUME

ED 462 402 TM 033 664

AUTHOR Reckase, Mark D.
TITLE Methodological Issues in Describing the Features and Qualities of Teacher Education Programs.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 2001-04-00
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Beginning Teachers; Elementary Education; *Evaluation Methods; Higher Education; Preservice Teacher Education; *Program Evaluation; *Research Methodology; Surveys

ABSTRACT

A complex research project has been conducted to determine the features and qualities of teacher education programs that are related to gains in student performance that occurred when a student was under the tutelage of a teacher from one of the teacher education programs. This paper describes the methodology and some of the results from work on producing good measurements of the multitude of variables that describe teacher preparation and professional judgment. Because of the large scope of the work, the discussion is confined to scale formation for three variables. The development team created the Beginning Teacher Preparation Survey to obtain information about a specific set of variables. Two efficacy scales, revised to one such scale, a professional development and support scale, and a scale measuring mathematics orientation provide examples of the sorts of scales developed for the survey. The confirmatory and exploratory reliability analysis process was performed on the majority of the hypothesized scales in the survey. A total of 26 scales have been defined and confirmed through the analyses. The lowest reliabilities have been in the 0.50 range and the highest in the upper 0.90 range. The distribution of scores for the scales were usually unimodal and often symmetric, but in some cases, the distributions were almost uniform. The result of the scaling methodology has been to produce a series of scales that are very sensitive to differences in beginning teacher's preparation and perspective. These scales can be used with high confidence to investigate the variables that lead to student learning in the classroom. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

M.D Reckase

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Methodological Issues in Describing the Features and Qualities of Teacher Education Programs[1]

Mark D. Reckase
Michigan State University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 462 402

The goal of this research project is to determine the features and qualities of teacher education programs that are related to those gains in student performance that can occurred while the student was under the tutelage of a teacher from one of the programs. Finding these relationships is a very complex process because many characteristics are needed to fully describe a teacher education program. Further, the functioning of the teacher education program also depends on the characteristics of the students entering the program and the match between the teacher education program and the characteristics of the school in which the teacher is functioning. The result is that a complex web of relationships is hypothesized to connect teacher education student characteristics to teacher education program characteristics to elementary school characteristics to the amount of student growth that can be attributed to the teacher.

Because of the complexity of the hypothesized model and the number of variables involved, it is unlikely that any one variable will be strongly related to gains in student performance. And that complexity does not consider the unreliability of scores and the number of student variables that are involved in

TM033664

---

[1] Paper presented at the annual meeting of the American Educational Research Association, Seattle, April 2001. This research was supported by a grant for the Office of Educational Research and Improvement.

BEST COPY AVAILABLE

the level of student performance (e.g., motivation, parental support, etc.). Detecting and modeling these relatively small relationships require fairly precise measurement of the relevant variables. This paper describes the methodology and some of the results from work on producing good measurements of the multitude of variables that describe teacher preparation and professional development. Because of the large scope of this work, it is not practical to describe the scale formation for every variable in the study. Instead, three variables will be given thorough coverage as examples of the process. Results for the other variables will be briefly summarized.

Conceptual Framework for Measuring the Variables

The goal of all measurement is to show true differences in the characteristic of interest. This goal is achieved by minimizing the error of measurement while at the same time using measurement tools (i.e., items) that are sensitive to the differences in the characteristic of interest. There are three general philosophical approaches to the development of measurement instruments: (1) domain sampling, (2) construct estimation, and (3) construction of an indicator.

The domain sampling approach to measurement is appropriate when the goal is to estimate the proportion of a large domain of behaviors that is exhibited by a person. A simple example is estimating the proportion of words in a dictionary that a person can spell correctly. The measurement is performed by randomly selecting a set of words from the dictionary and asking the person to

spell them. The proportion of the sample of words that are spelled correctly is used as an estimate of the proportion of the full domain of words that can be spelled correctly.

The construct estimation approach is appropriate when there is a hypothetical continuum of skills, attitudes, etc., and the goal is to locate a person on the continuum. A common example of a hypothetical construct is verbal aptitude. Persons are placed on the continuum for the construct using their responses to a variety of verbal tasks.

The construction of an indicator is appropriate when it is expected that a collection of characteristics is likely to be predictive of an outcome, but when no single domain or continuum is hypothesized to exist. For example, a constructed indicator of the likelihood of completing a college degree is financial support plus good grades plus stable social environment plus reasonable health. For each student, a yes/no response can be obtained for each component of the indicator. A score of four indicates that all four components are present and it is hypothesized that a person with a four would have a high probability of completing a degree program. A score of zero indicates the person is likely to drop out before getting a degree. There is no domain of skills or hypothetical construct behind this indicator. It is only a constructed index that the developer believes will be related to the criterion behavior.

The Beginning Teacher Preparation Survey includes variables of a variety of types. In the next section, three examples are discussed in detail, and the measurement philosophy behind each variable is described.

Scale Development for the Beginning Teacher Preparation Survey

The Beginning Teacher Preparation Survey was developed by a team of researchers with expertise in a wide variety of educational areas from curriculum to pedagogy to educational testing. Many of the survey items were selected from previous work on the evaluation of teacher preparation programs. Additional items were produced by the development team to tap variables identified in the teacher development literature.

After the pool of survey items was produced, the items were pilot tested on a small sample of graduate students in a college of education to identify items that did not function properly. The statements may have been unclear, the terminology might not have been familiar, or there may have been awkward phrasing. All comments from the pilot test sample were reviewed to identify items that should be deleted from the pool or revised.

After review and revision, the item pool was judged to be too large to administer in a reasonable period of time. Because of concerns that the response rate to a mail-out survey of such length would be very low, a subset of the full pool of items was selected that the developers believed could be administered in an hour or less. The selection process had the goal of maintaining the coverage of the desired variables with high quality scales.

Redundant items were eliminated from consideration, and items were identified that had clear connections to the variables. The resulting survey still had over 400 items, yielding very rich data on teacher preparation.

The challenge to the methodological portion of this study was to develop a set of highly reliable measures from the full set of responses that captured information about the desired set of variables. Individual items are unreliable so they are unlikely to be useful for detecting subtle relationships in the data. Therefore, items were combined into scales to obtain scores that are more reliable. It was also desirable to have scores that were roughly normally distributed to support the assumptions of future statistical analyses.

A four-step scale development process was implemented to achieve the goals of producing reliable and valid scales with good statistical properties. First, the survey development team identified sets of items that they believed would logically fit together to form scales. These sets of items were identified from a review of previous research and the expert judgement of the development team. The second step was to perform confirmatory analyses to determine if the empirically defined scales were supported by the relationships in the empirical data. For support to be present, the teachers in the sample had to vary on the hypothesized construct or domain, and the items had to be sensitive to differences on the construct or domain. If empirical data supported the scale, the reliability of the scale was estimated and the score distribution was computed. This was the third step in the process.

If the empirical data did not support the hypothesized scales, exploratory analyses were conducted to develop new hypothesized variables. The results of these analyses were shared with the other members of the development team so they could determine whether the scales were supported by the research literature. If there was support, new scales were constructed and reliability and score distributions were estimated. In all cases, the goal was to create scales that were supported by prior research and that had good technical quality. In no case was a scale produced solely based on statistical analyses. The process of scale development is described in the next section for several of the scales.

The Beginning Teacher Preparation Scales

The development team created the Beginning Teacher Preparation Survey to obtain information about a specific set of variables. For each of the variables that was the target for the survey, the development team identified the set of survey items that they believed would form a scale. Table 1 provides a list of the hypothesized scales and the items that were thought to be related to the scale. Several of the scales will now be discussed in detail.

Insert Table 1 about here

The Efficacy Scales

The development team hypothesized that two efficacy scales would be supported by the survey data: general efficacy and personal efficacy. These

scales follow the hypothetical construct conception of scale development because it was expected that teachers could be placed along a continuum from low to high efficacy for affecting the performance of students. A typical item on the general efficacy scale is "Teachers can do little to overcome the effects of students' lack of motivation." Teachers responded to this item using a rating scale from "strongly disagree" to "strongly agree" and the ratings were reverse scored so that "strongly disagree" indicated positive general efficacy. Overall, general efficacy was to be measured by reactions to statements about teachers' abilities to make a difference in students' performance.

A typical item on the personal efficacy scale was "Improvement in my knowledge and skills will result in improvement in my students' academic performance." The ratings for this item were scored in the positive direction. Personal efficacy items relate to things a specific teacher can do rather than what teachers in general can do.

The confirmatory analysis of the 13 items in the two hypothesized scales did not support two separate scales. The items from the two different scales correlated more highly with each other than items within the scales. To get a better understanding of the efficacy scales, a factor analysis with oblique rotation was performed. That analysis supported a single efficacy scale using the majority of the items, but it also identified some minor other factors that had to do with working with second language learners and the responsibilities of teachers. Because of the small number of items related to these other factors,

and the fact that the focus of the study was on the general student population, no attempt was made to produce separate subscales using the items. Instead, only the dominant subscale was retained for the study.

The final Efficacy scale consisted of 7 items. A scale score was computed by summing the item ratings after orienting the ratings so that positive scores meant high efficacy. The resulting scale scores ranged from 9 to 35. The mean, standard deviation, and coefficient alpha reliability for the scale are given in Figure 1. The reliability of .66 is only moderate, but it is in a range that is sufficient for research applications. Generally, the analyses support that the measure can be used to order teachers along an efficacy continuum.

Insert Figure 1 about here

Professional Development and Support

Two sets of items were originally hypothesized to be measures of Professional Development and Support. The first set of items was a listing of a sampling of types of support from a domain of possible types of such support. Examples of support options included "reduced teaching schedule" and "extra classroom assistance." Teachers responded either "yes" they received such support, or "no" they did not. These items follow the domain sampling conception of scale development.

The second set of items was related to experience working with a mentor. Teachers rated the frequency of types of activities from "never" to "weekly," and the value of the activities from "not at all valuable" to "very valuable." These items follow a scale formation philosophy that is a cross between sampling from a domain of activities and forming a hypothetical construct called mentoring effect.

Given the variety of items in the professional development support scale, it is probably not surprising that a single scale was not confirmed by the analysis. Exploratory analyses indicated that the mentoring frequency and value were two different variables. Further, the domain sampling of support activities was not related to mentoring activities. In fact, the yes/no responses to the supporting activities were not very highly related and did not seem to merit a scale. The results of these analyses and discussions with the development team were that two scales were formed -- mentoring frequency and mentoring value – by summing the ratings of the items for those sets of items. The score distributions, means, and standard deviations, and coefficient alpha reliabilities are given in Figure 2. The two scales are correlated .88 so a total mentoring experience was also developed. Note that these variables are not normally distributed and they may have to be transformed to meet the assumptions of some statistical analysis procedures.

Insert Figure 2 about here

Mathematics Orientation

A set of items was included in the survey that describe a teacher's beliefs about teaching mathematics. The items require ratings of statements like "The main job of a teacher is to transmit knowledge and content of mathematics" from" from "strongly disagree" to "strongly agree." When thinking about these items if is important to remember that elementary school teachers are responding to the items and they may have a different perspective than secondary teachers.

The confirmatory analysis of the items did not support a single construct for mathematics orientation. Exploratory analyses suggested three scales. Interestingly, development team indicated that they had intended to have two different types of mathematics orientation items when the scale was developed. One type of orientation item considered attitudes toward the mathematics reform movement. The second type of item considered attitudes toward traditional ways of teaching mathematics. The exploratory analysis identified these two dimensions in the response data, and also a third dimension. The third set of items indicated an approach mathematics instruction that stresses making sense of the mathematics and the students' learning style. That is teachers were attempting to understand each student's learning style so that they could bring about student understanding of the mathematics. Based on the analyses and the reactions from the development team, three scales were developed for

mathematics orientation. Although they were originally thought of as reform orientation, traditional orientation, and adaptability, more value neutral titles are Group and Project Activities, Drill and Lecture Activities, and Sense-Making Activities.

The distributions of scores on these scale, the means, standard deviations, and reliabilities are shown in Figure 3. Note that the distributions of the three scales are quite different. Many of the teachers indicated that they frequently used drill and lecture activities. Group and project activities were more normally distributed. Sense-making activities were exhibited quite frequently, but there was still quite a bit of variation in the amount of sense making.

Insert Figure 3 about here

Summary and Conclusions

The confirmatory, exploratory, reliability analysis process has been performed on the majority of the hypothesized scales in the Beginning Teacher Preparation Survey. A total of 26 scales have been defined and confirmed through the analyses. The scales are based on well-defined measurement philosophies; usually domain sampling or hypothetical construct estimation. In rare cases, constructed scales were created. These were predominantly used to indicate the type of preparation and license that had been obtained. The scales

all have moderate to high reliabilities. The lowest reliabilities were in the .50s and the highest in the high .90s. The distribution of scores for the scales was usually unimodal, and often symmetric. However, in some cases such as mentoring, the distributions were almost uniform.

The overall result of this scaling methodology has been to produce a series of scales that are very sensitive to differences in beginning teachers' preparation and perspective. These scales can be used with high confidence to investigate the variables that lead to student learning in the classroom.

Table 1
Hypothesized Scales
Beginning Teacher Preparation Survey

Teacher Preparation/Induction Scales

	Structural Factors	*Items*
1	Licensure Route	I-3, 4, 5
2	Program Type	I-6
3	Coherence [within program]	B1-4, 24, 29-32
4	Faculty Characteristics	B1, 8, 23-26, 28, 33
5a	Field Experiences (PDS)	B9, 12-16, 27, 33, I9
5b	Theory-Practice Relation	B5, 8, 16, 23-25, 28—this factor might be eliminated since each of these items is in another factor
6	Candidate Assessment	B17-22
7	Alignment [between preservice & teaching assignment]	A6, B5, 6, 16, 34-37, [I-1&4 and A5 comparison]
	Conceptual Factors	
8	Subject Matter Preparation	B7, 10, 11, I1, 2
9	Pedagogical Preparation	C1-14, 21
10	Diversity Preparation	C15-23
11	Prep for Reading Instruction	C24-38
12	Prep for Math Instruction	C39-51
13	Prep for Student Assessment	C52-59
	Overall Factors	
14	Program Quality	C60
15	Program Impact	B38-43
	Induction (PD) Factors	
16	PD Support	G4, 5
17	PD Focus & Quality	G1
18	PD Characteristics (Form)	G2, 3
19	PD Impact	G6

Table 1 (Continued)

	Teacher Belief/Knowledge/Practice Scales	
20	General Efficacy	D1, 4, 7, 9, 10, 11, 12
21	Personal Efficacy	D2, 3, 5, 6, 8, 13
22	Literacy Knowledge	E4, I-1, 2, 17j-l
23	Literacy Orientation	E1
24	Mathematics Orientation	F1
25	Mathematics Knowledge	F5, I –1, 2, 17m-n
26	Literacy Materials	E3
27	Literacy Activities	E2
28	Mathematics Materials	F4
29	Mathematics Activities	F2, 3
30	Pedagogical Knowledge	I-17g-i
	Control/Sorting Scales	
31	Individual	A4-6, I- 7, 10-17a-f
32	Classroom	H1-3, I-8
33	School	H4

Figure 1

Observed Distribution of the Efficacy Variable

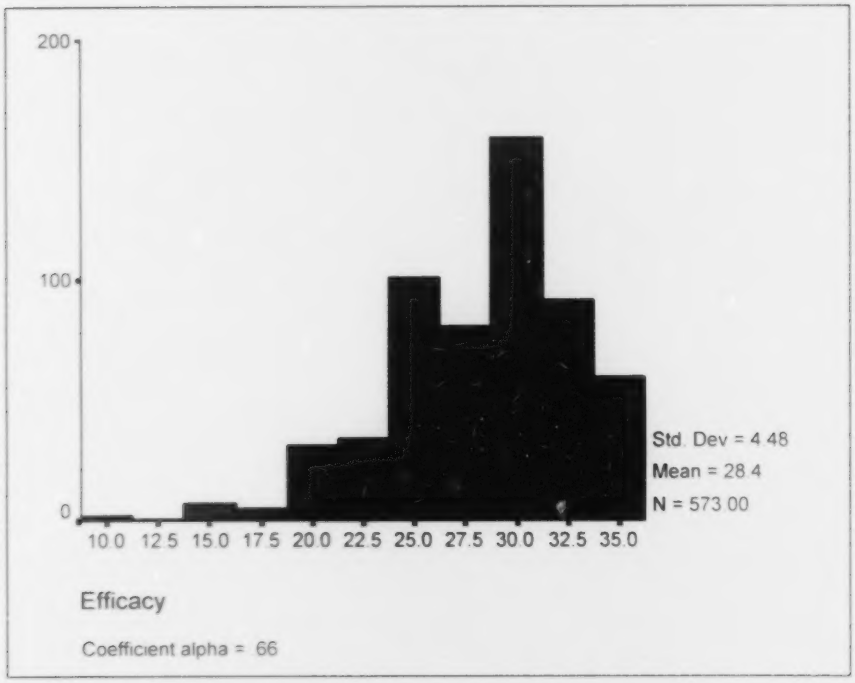
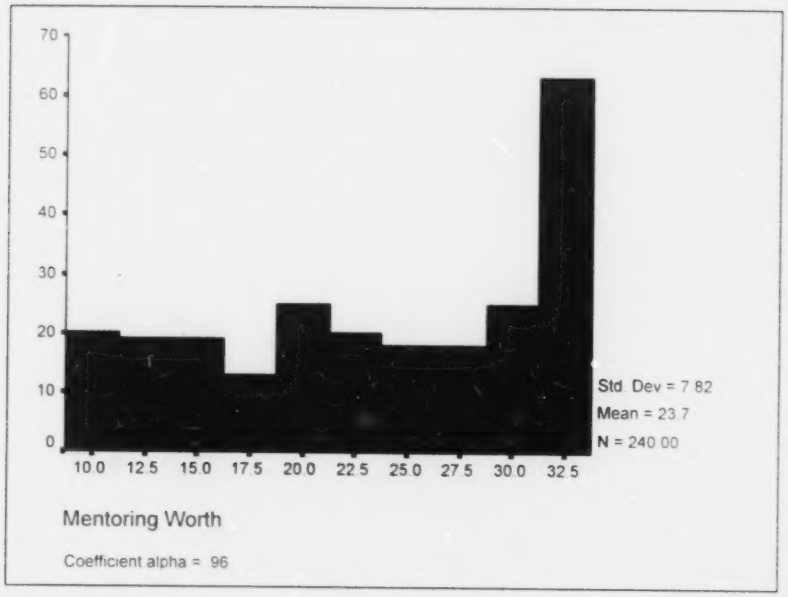
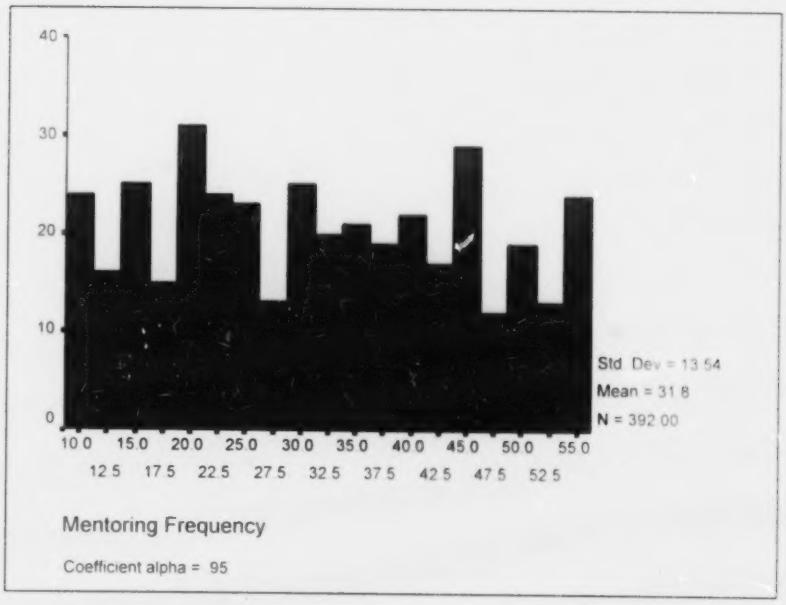


Figure 2

Mentoring Variables





BEST COPY AVAILABLE

Figure 2 (Continued)

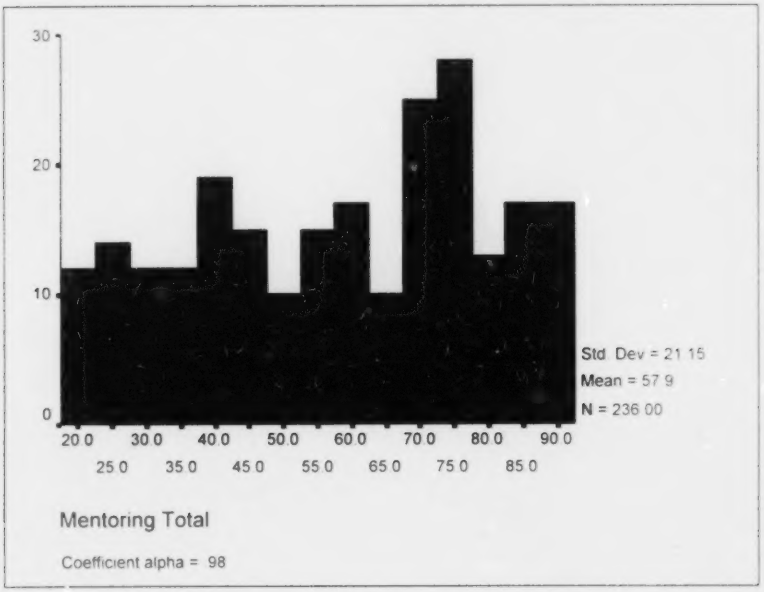
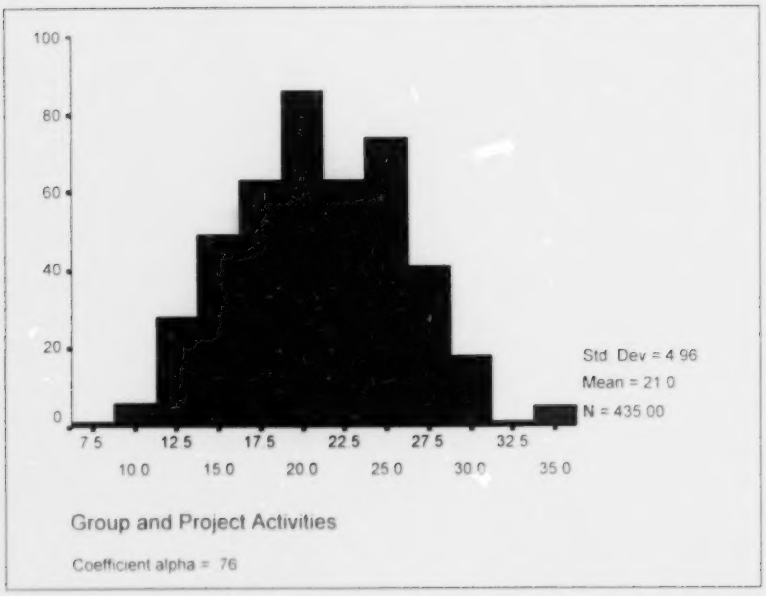


Figure 3

Mathematics Orientation Scales



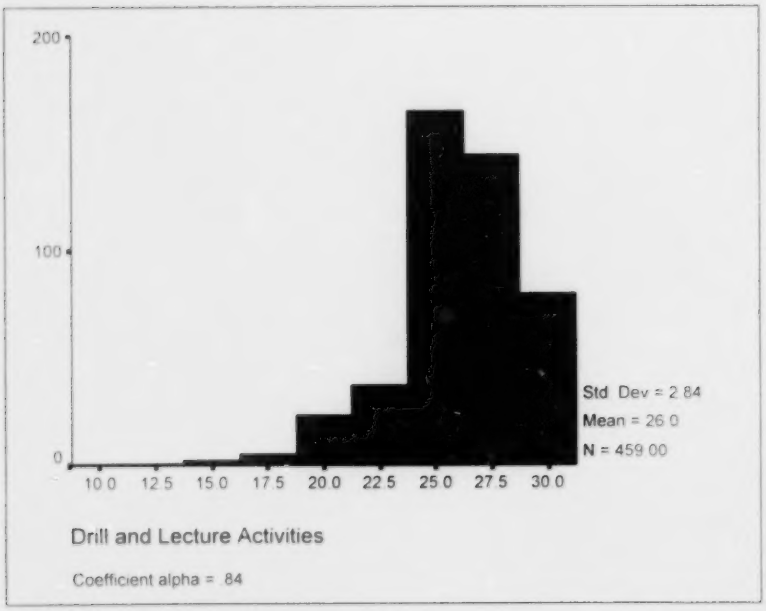


Figure 3 (Continued)